Proyecto Final - Probabilidad y Estadística Aplicada

Grupo:
Estefany Clara, Gonzalo Paz, Juan Pérez y Lucas Cordero

Introducción:

NOTES __/__/__

Sean $x_1, x_2, \dots x_n$ iid

Con distribución normal

con $E(x_1) = \mu$ → desconocido. ?

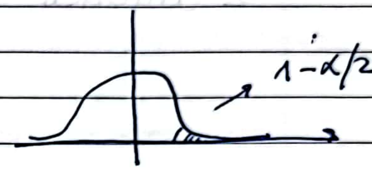
$Var(x_1) = \sigma^2$ conocida.

Varianza (la uso en la formula final).

$$I_\alpha = \left[\bar{x}_n - \frac{\sigma}{\sqrt{n}} z_{\alpha/2},\ x_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right]$$

Intervalo de confianza

Varianza conocida.



$$P(\mu \in I_\alpha) = 1 - \alpha$$

Ballotage
de que me sirve que el candidato 48% +-5%. porque no se si el candidato va a obtener el 51%

Varianza desconocida

Sea $x_1, x_2, \dots, x_n$ iid con distribución normal con

$E(x_1) = \mu$, $Var(x_1) = \sigma^2$ desconocidas.

Teoria: la variable aleatoria

$$\frac{\sqrt{n}\,(\bar{x}_n - \mu)}{S_n} \qquad S_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$$

NOTES ___/___/___

Intervalo de confianza.

medicina muy poca %.

parametrización.

2 variables

st.t intervalo
x lo de la clase
foto pizarron.

$df = len(tabb) - 1$

$\underbrace{t_{\alpha/2, n-1}}_{\text{python.}}$

loc = media empírica.

NOTES ___/___/___

Test de hipotesis. que es un test de hipotesis?

$X_1, X_2, \ldots, X_n$ muestra iid con distribucion desconocida.
¿la muestra tiene distribución normal? ¿si o no?

¿la media poblacional (μ) es mayor a cierto valor? (μ_0)

¿como con pocos datos?

completa
V/F

Test de hipotesis: Un contraste de hipotesis es una herramienta estadistica para intentar decidir entre 2 hipotesis que llamaremos H_0 y H_1.

Ejemplo: se tira un dado 100 veces.

H_0 El dado esta equilibrado.
H_1 No H_0

Decision en un test de hipotesis:

Rechazar H_0 (en favor de H_1)
realidad
H_1 Verdadera ✓
H_0 verdadera. Error tipo 1 X
α

No rechazar H_0
realidad
H_1 verdadera X → error tipo 2
H_0 verdadera ✓
P(error tipo 2) = β

NOTES ___/___/___

Aclarar 0

En un preparado alimenticio infantil se especifica que el contenido medio de proteínas es de (42%) μ. Tratamos de comprobar esta especificación, para ello tomamos 10 de estos preparados alimenticios para determinar su contenido de proteínas. Se obtiene un promedio muestral $\bar{X}_{10} = 40\%$, y una desviación estándar muestral de $S_{10} = 3,5\%$.

Diseñar un test de hipótesis que decida si la especificación es correcta.

(1) Cosa escribir que test vamos a plantear.

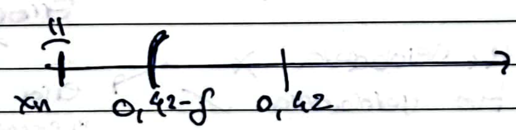
Planteamos el siguiente test de hipótesis.

$$\begin{cases} H_0: \mu = 0,42 \\ H_1: \mu < 0,42 \end{cases}$$

donde $\mu = E(x_1)$

$x_1, x_2, \ldots, x_n \sim N(\mu, \sigma^2)$ desconocidos $n = 10$

Propongo el siguiente criterio de decisión.



Calculamos $0,42-\delta$ si es mucho menor el promedio ... significa que todo real si queda x abierto

NOTES ___/___/___

la idea hallar un número $\delta > 0$ / si $\bar{x}_n < 0{,}42 - \delta$ (H_1) elegimos H_1 y si está si $0{,}42 - \delta \leq \bar{x}_n < 0{,}42$ elegimos H_0

Hallaremos δ para que con ese criterio de decisión $\alpha = P(\text{error tipo } 1) = 0{,}05$.

$$\alpha = P(\text{Elegir } H_1 \mid H_0 \text{ verdadero})$$

cuando

$$\Downarrow \qquad \Downarrow$$

$$\alpha = P(\bar{x}_n < 0{,}42 - \delta \mid \mu = 0{,}42)$$

$$\alpha = P(\bar{x}_n - 0{,}42 < -\delta \mid \mu = 0{,}42)$$

$$\alpha = P\left(\frac{\bar{x}_n - 0{,}4}{s_n} < \frac{-\delta}{s_n} \;\middle|\; \mu = 0{,}42\right)$$ entre los dos lados dividido entre s_n.

T_{n-1} T student.

$$\alpha = P\left(\sqrt{n} \cdot \frac{(\bar{x}_n - 0{,}42)}{s_n} < \frac{\sqrt{n}(-\delta)}{s_n} \;\middle|\; \mu = 0{,}42\right)$$

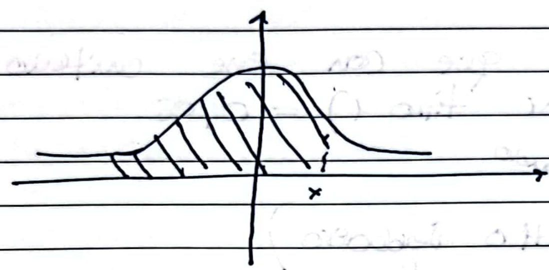
xq? esto! Teorema: la variable aleatoria

$$\left(\frac{\sqrt{n}(x_n - \mu)}{s_n}\right) \sim T_{n-1}$$

→ es lo mismo cambia μ ~~[illegible]~~ a grado d distribució.

NOTES __/__/__

~~Fro~~ import scpy.stats as st.



funciones ① st.t.cdf para distribucion acumulada de T. students.
② st.t.~~cdf~~ppf función cuantil (inversa de cdf)

$$\text{st.t.cdf}(x, df=9) = P(T_9 \leq x) \quad \Big\}\ \text{lado izquierdo}$$

$$\alpha = P\left(t_9 < \frac{\sqrt{n}\,(-\delta)}{s_n}\right)$$

$$\alpha = \text{st.t.cdf}\left(\sqrt{n} \cdot \frac{(-\delta)}{s_n}\right)$$

$$\text{st.t.ppf}(\alpha) = \sqrt{n} \cdot \frac{(-\delta)}{s_n}$$

$$-\frac{s_n\ \text{st.t.ppf}(\alpha)}{\sqrt{n}} * {}^{\text{grados libertad}} = \delta$$

$$\text{st.t.ppf}(0{,}05,\ df=9) = -1{,}833$$

NOTES

$\delta = \frac{-0{,}035 \cdot (-1{,}833)}{\sqrt{10}} \approx \boxed{0{,}0203.}$

Criterio : Si $\bar{x}_n < 0{,}42 - \delta$ decidimos H_1.
si $0{,}42 - \delta \leq \bar{x}_n < 0{,}42$ decidimos H_0.

tener en cuenta decisión

El test atribuye que es un error.

Terminologia para el projecto

Nivel: → H_0 se contuvo en 0,7 %

$X_1, X_2, \dots X_n \sim N(\mu, \sigma^2)$ iid.

$\begin{cases} H_0 = \mu = \mu_0 \\ H_1 = \mu < \mu_0. \end{cases}$

Test de una cola

st.st.ppf (q, df = n-1)

Decidir H_1: $\bar{x}_n < \mu_0 - \frac{S_n \, t_{1-\alpha, n-1}}{\sqrt{n}}.$

Si quisieramos hacer otro caso.

Test de una cola, unilateral.

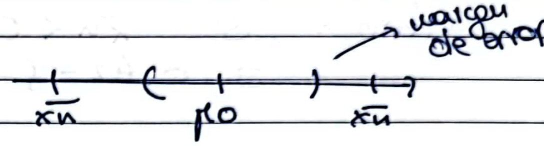
$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 = \mu > \mu_0. \end{cases}$ $\bar{x}_n > \mu_0 + \frac{S_n \cdot t_{1-\alpha, n-1}}{\sqrt{n}}$

NOTES ___/___/___

Bilateral el del proyecto

para los dos lados

$$\begin{cases} H_0 & \mu = \mu_0 \\ H_1 : & \mu \neq \mu_0. \end{cases}$$



Decidimos H_1 si $\bar{x}_n > \mu_0 + \delta$ o si $\bar{x}_n < \mu_0 - \delta$

Con $\delta = \dfrac{S_n \cdot t_{1-\frac{\alpha}{2},\, n-1}}{\sqrt{n}}$

st.t.ppf

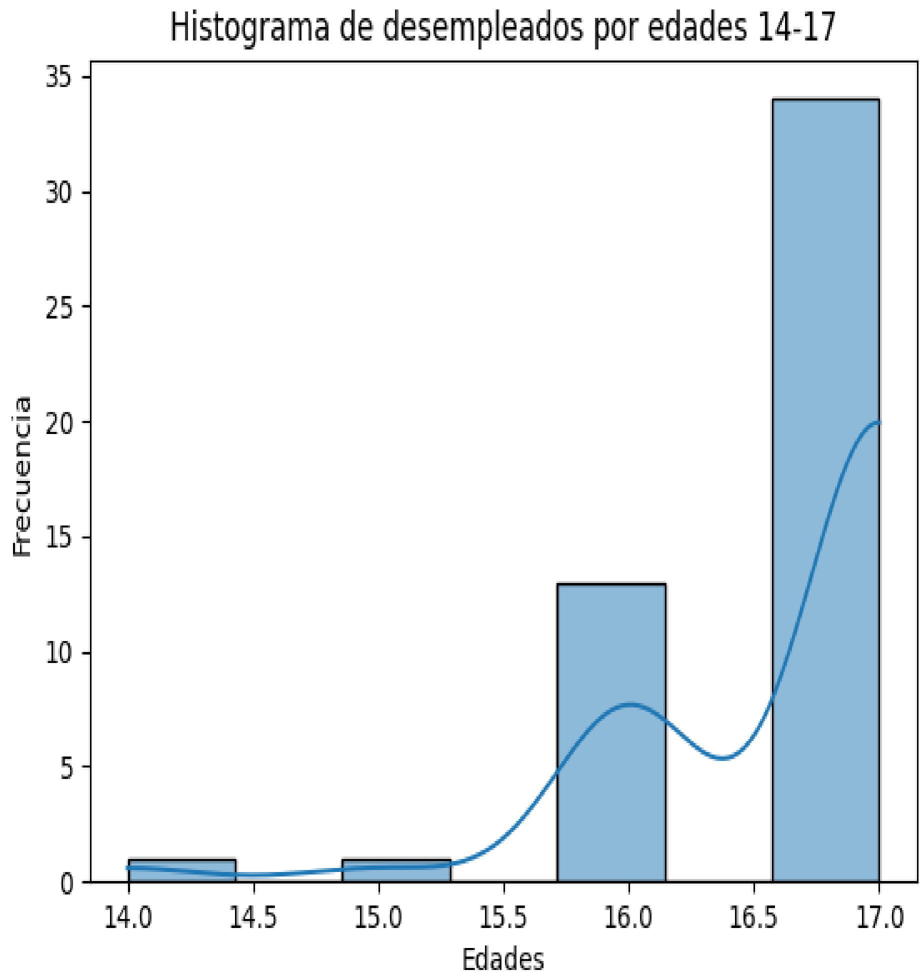
$(\frac{\alpha}{2}$, df = n-1

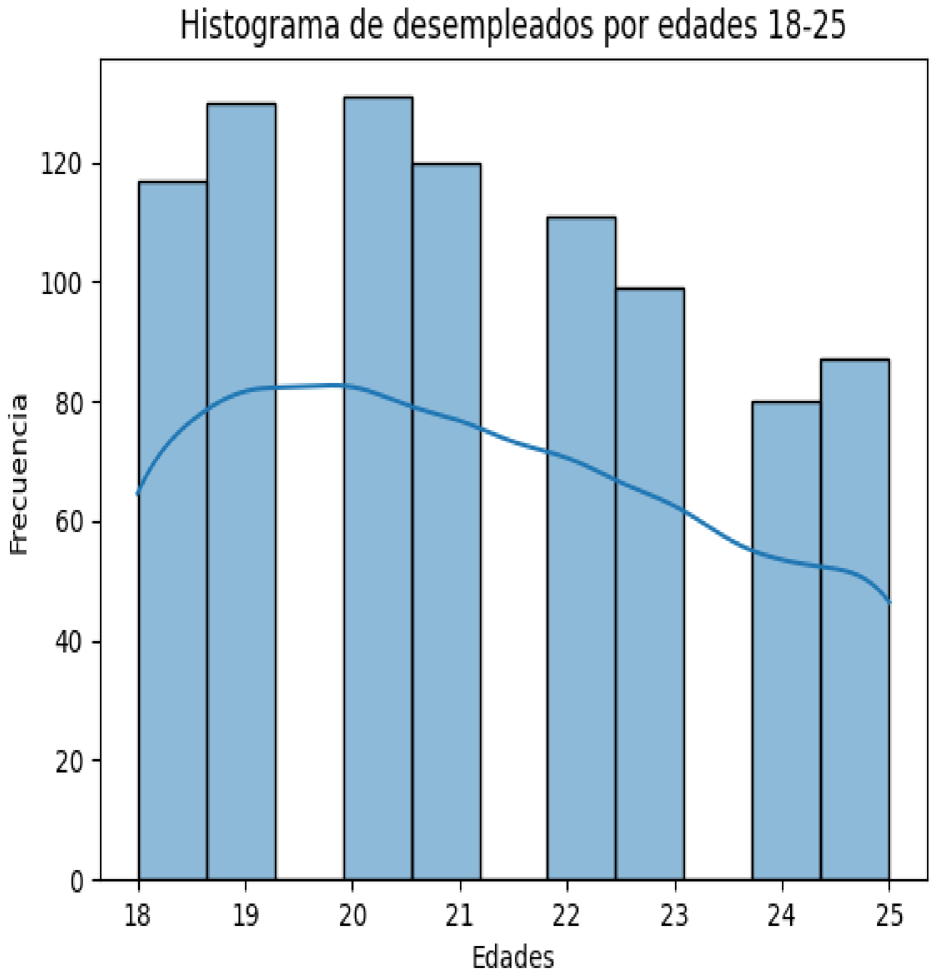
Anexo al informe:

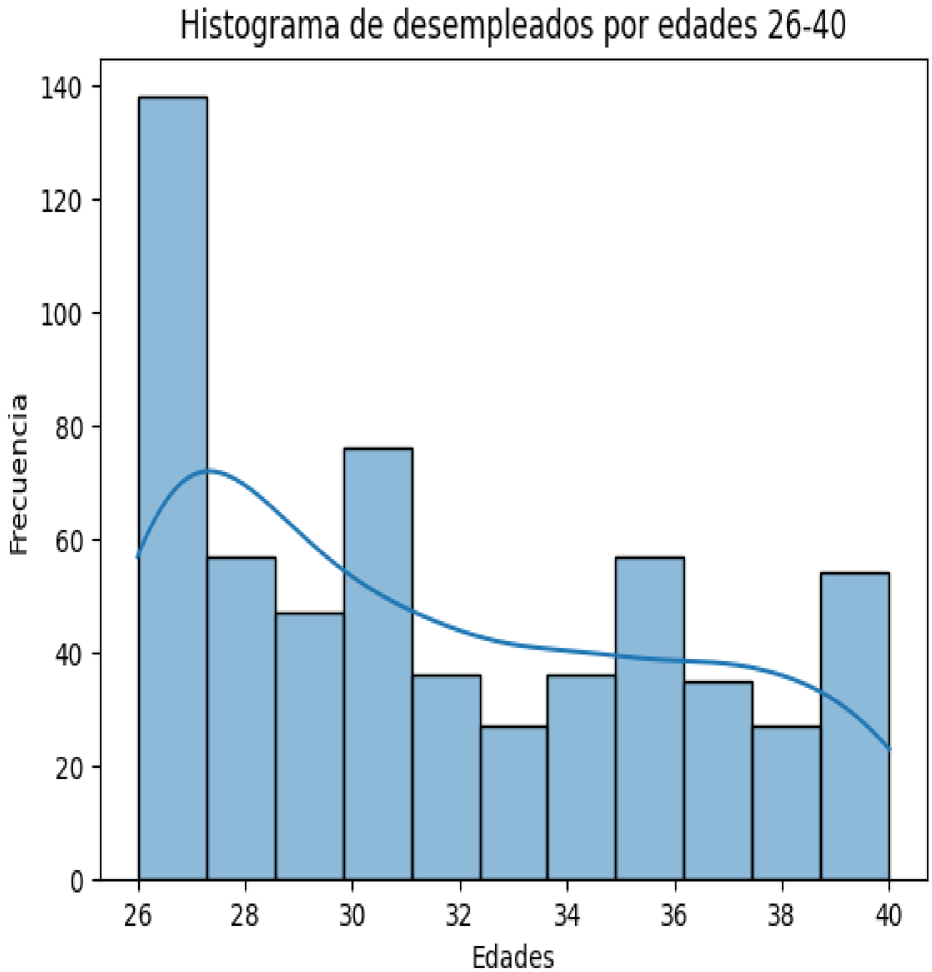
Parte A - DESEMPLEADOS

La cantidad de desempleados presentados en la muestra son: 2050
La cantidad de personas en la muestra son: 46522
1- a- La taza de desempleo es: 0.04406517346631701

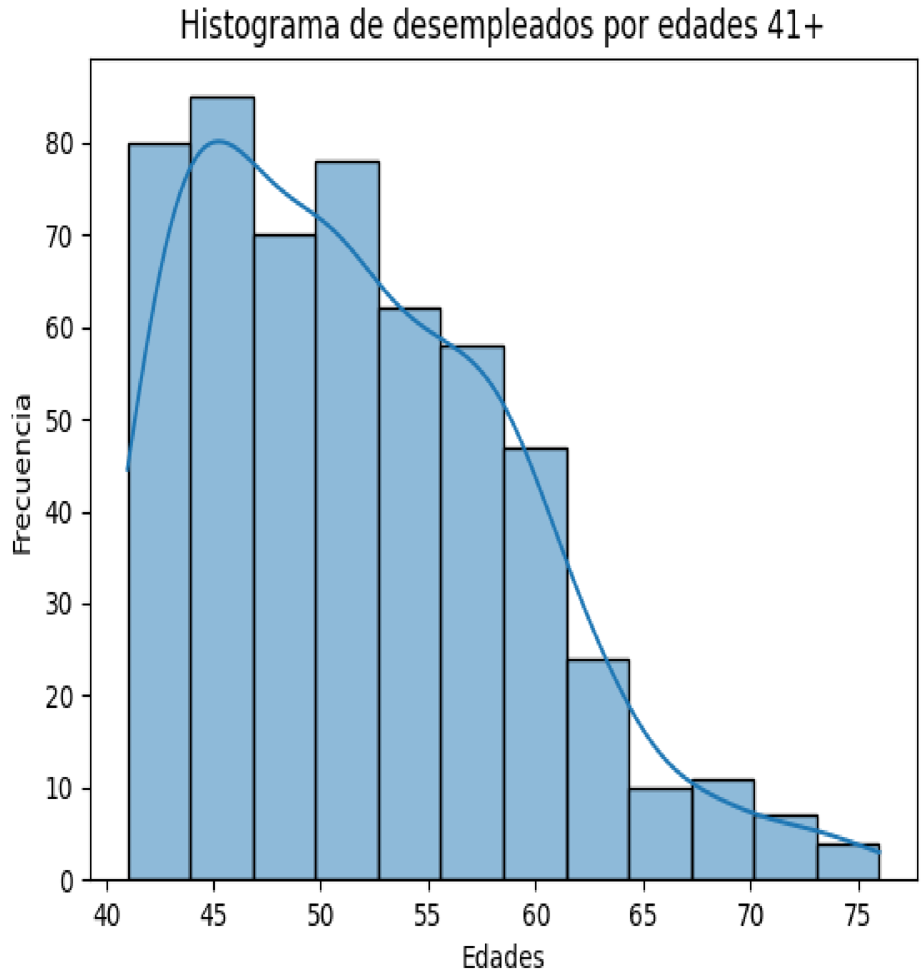
Parte B - HISTOGRAMA DESEMPLEADOS



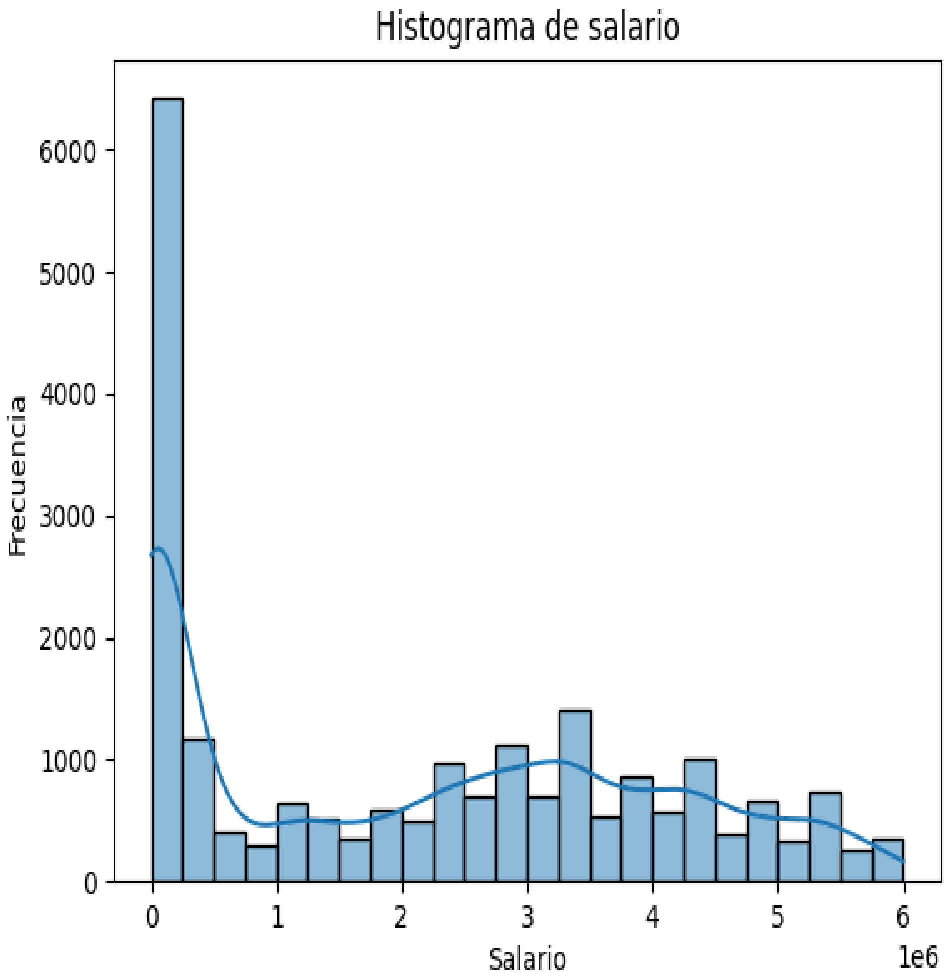

Histograma de desempleados por edades 18-25
Frecuencia
0
20
40
60
80
100
120
18
19
20
21
22
23
24
25
Edades



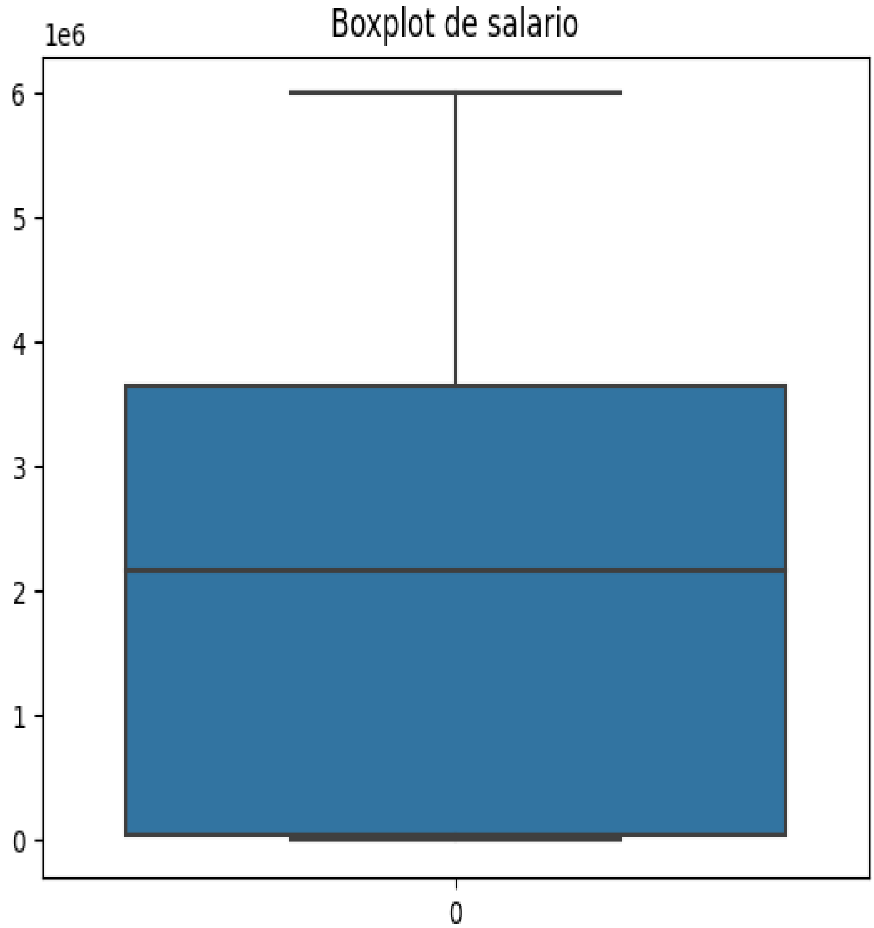
Histograma de desempleados por edades 26-40
Frecuencia
0
20
40
60
80
100
120
140
26
28
30
32
34
36
38
40
Edades


Histograma de desempleados por edades 41+
Frecuencia
0
10
20
30
40
50
60
70
80
40
45
50
55
60
65
70
75
Edades

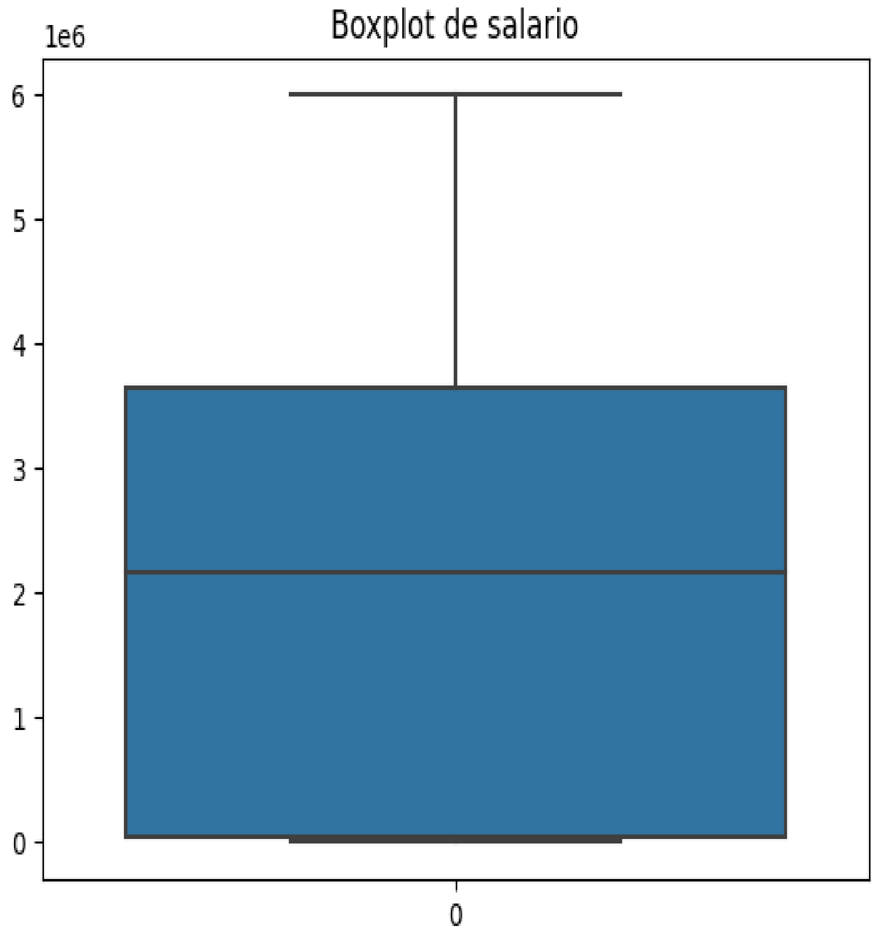
Parte 2 - SALARIO
2) a. Elaborar un histograma de la variable salario



2) b. Elaborar un Boxplot para toda la muestra



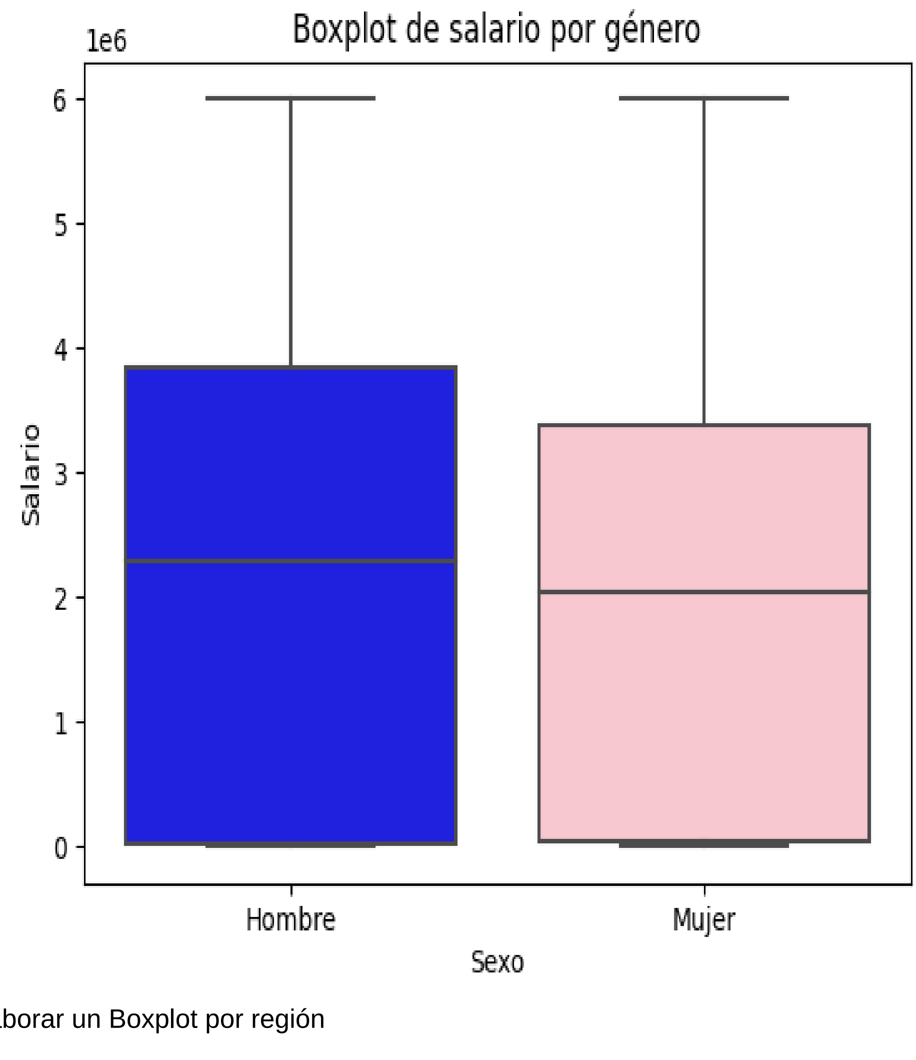
ii. Corrección de valores atípicos



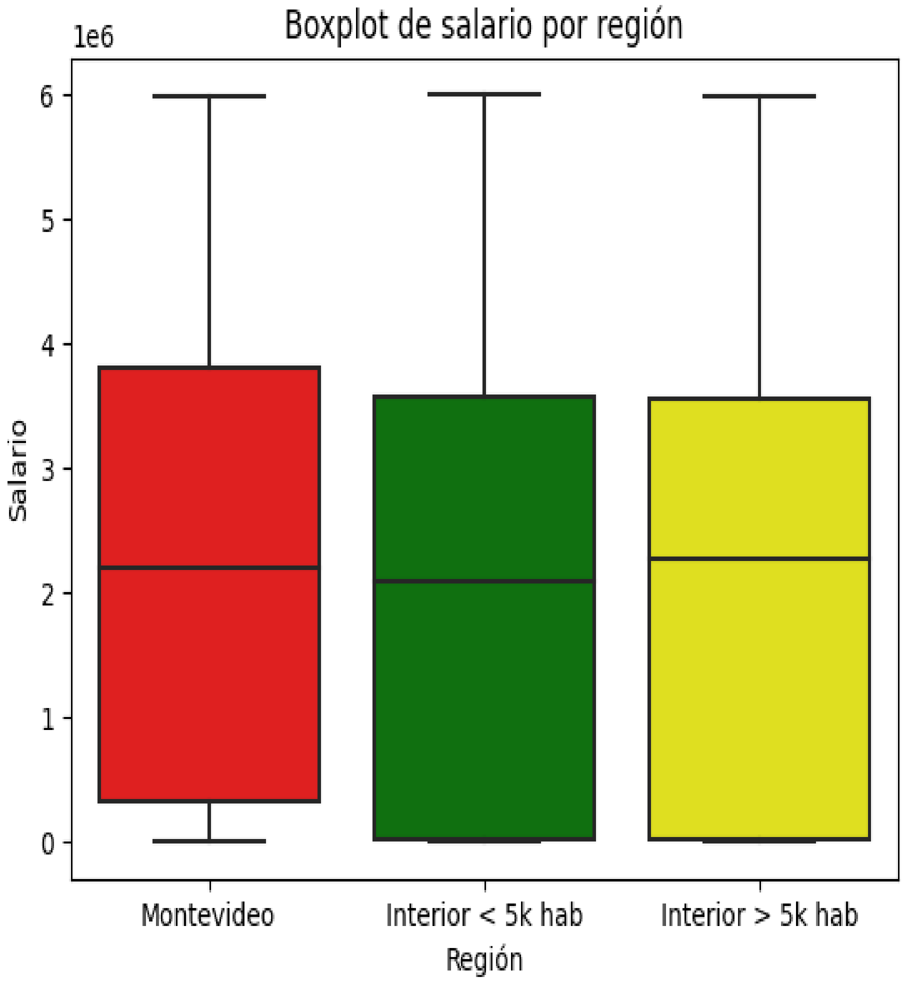
2) c. Media: 2120986.549251783, Mediana: 2165047.0, Moda: ModeResult(mode=array([20000], dtype=int64), count=array([420]))

Valores de estudio del Boxplot de Salarios
2) d. Q1: 2165047.0, Q3: 35000.0, Mínimo: 3655501.5, Máximo: 50, RI: 5999367, BII: 3620501.5, BSI: -5395752.25

2) e. i. Elaborar un Boxplot por género (masculino y femenino)



2) e. ii. Elaborar un Boxplot por región



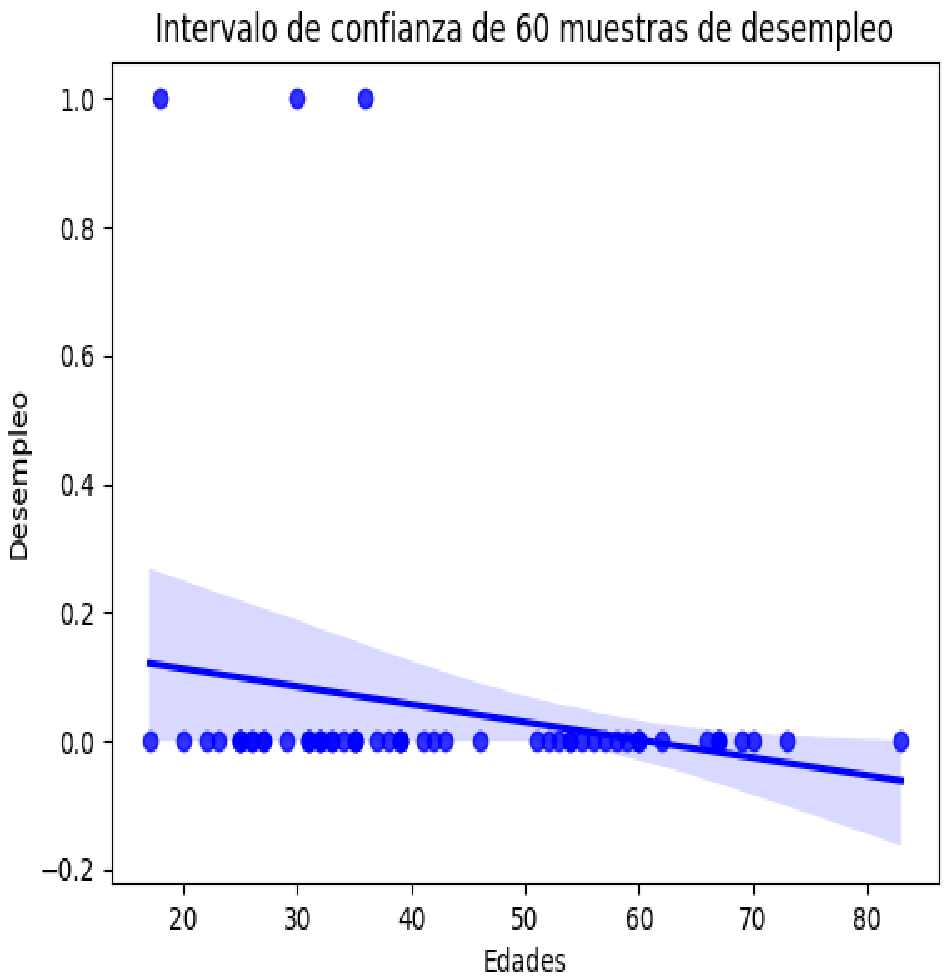
Boxplot de salario por región
1e6
6
5
4
3
2
1
0
Salario
Montevideo
Interior < 5k hab
Interior > 5k hab
Región

PARTE B)- ESTIMACIÓN

1) Estimar el desempleo del total de la población

El total estimado de desempleados es: 132564

2) Elabora intervalo de confianza con 95% de certeza para la variable desempleo.



El intervalo de confianza para la variable Desempleo es: (0.04220009652080195, 0.045930250411832074)
La muestra arroja que existe un intervalo de certeza del 95% que el desempleo es correcto según la muestra.

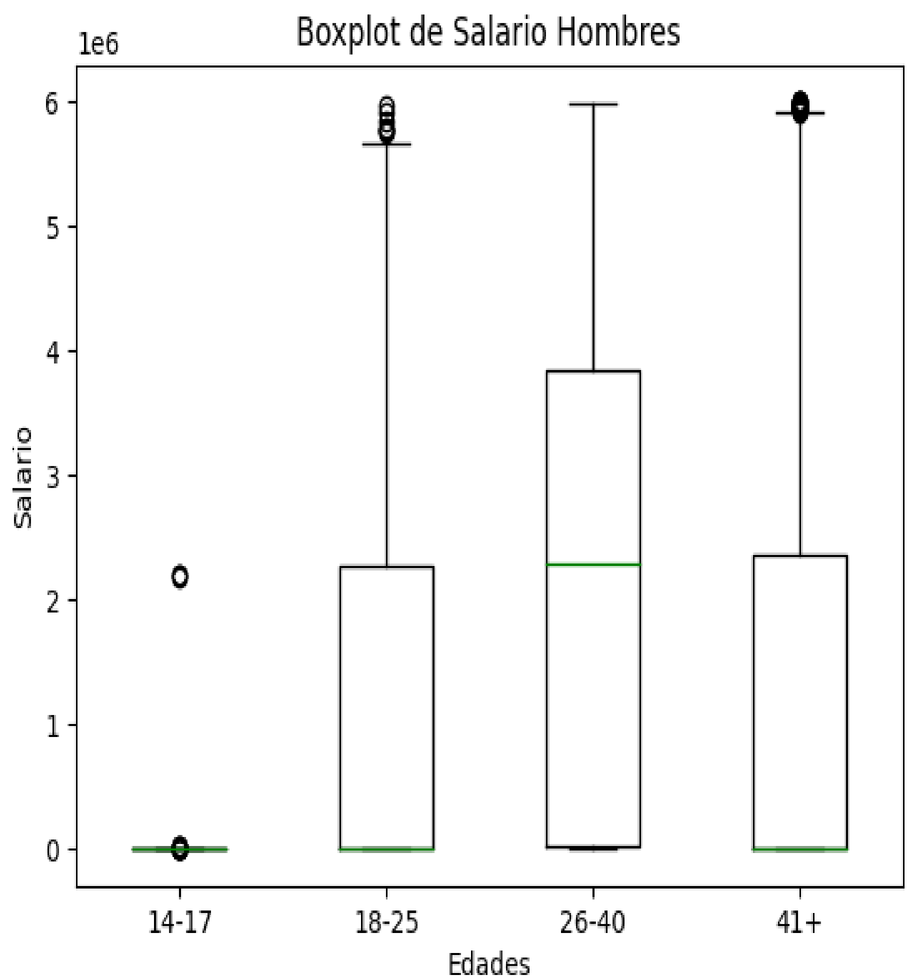
C) PRUEBA DE HIÓTESIS

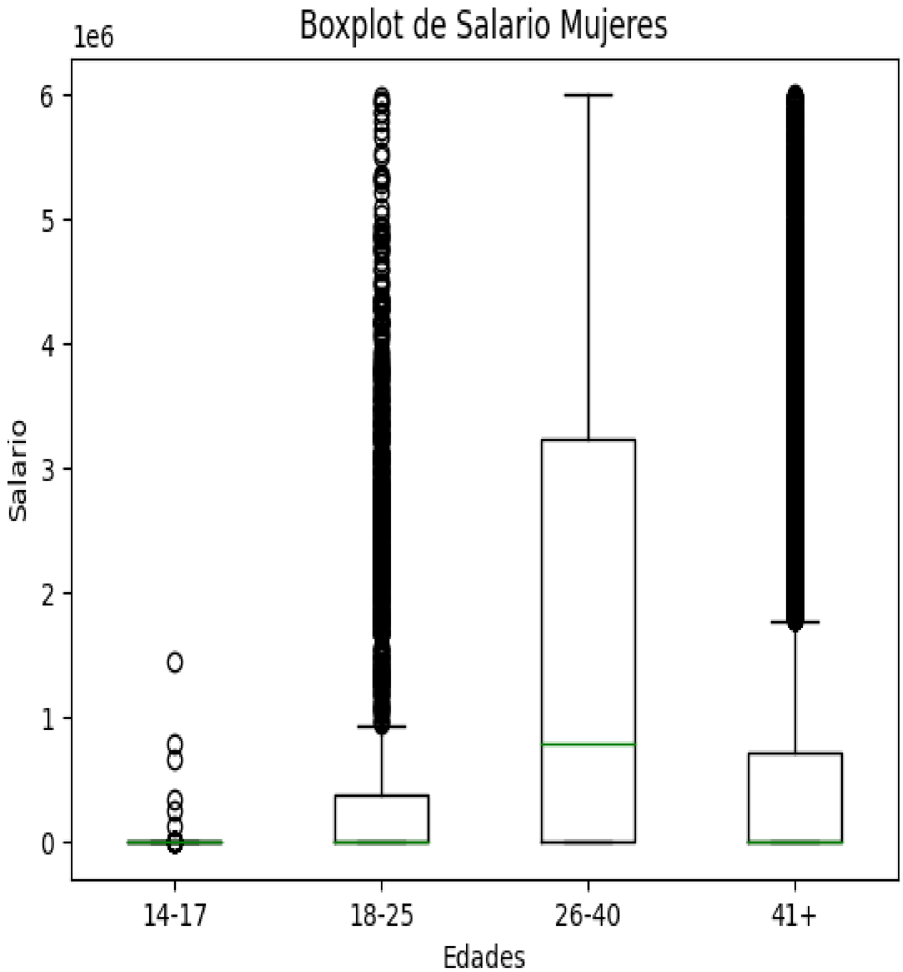
1) DESEMPLEO:

Se pide verificar si es correcto decir que la tasa de desempleo aumento desde 2021.
t_static: -6.2369270791397895
p_value: 4.5007738606764015e-10
No se rechaza la hipótesis nula.

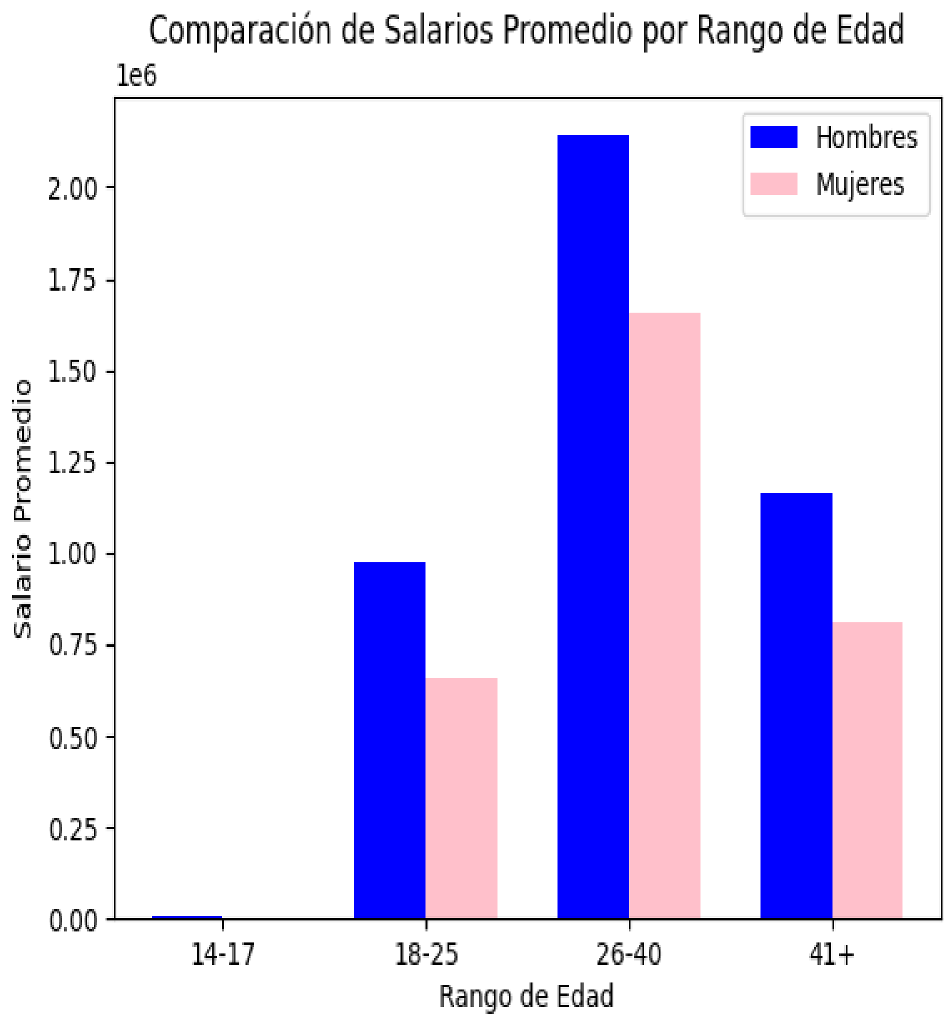
2) SALARIO:

Se pide examinar la muestra con certeza de 99%, para verificar si el salario promedio entre los hombres y las mujeres es el mismo.





Boxplot de Salario Mujeres
1e6
6
5
4
3
2
1
0
Salario
14-17
18-25
26-40
41+
Edades


Comparación de Salarios Promedio por Rango de Edad
1e6
Hombres
Mujeres
Salario Promedio
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
14-17
18-25
26-40
41+
Rango de Edad

Conclusiones:

Como grupo creemos que a este trabajo por ser el de final de Unidad, nos llevo más tiempo poder realizarlo. Debido a varios factores, como la complejidad del mismo, así como también que en algunas partes de la letra nos hubiera gustado tener más ideas de lo que se nos estaba pidiendo, ya que no quedaba muy claro.
Por suerte para nosotros los apuntes tomados en clase, las respuestas de los profes a dudas de los compañeros y algunas propias, nos ayudaron a entender un poco mejor que es lo que se quería conseguir con este trabajo.
También queremos destacar que poder ver los resultados de los conceptos aprendidos en clase, en un trabajo final, nos ayuda a entender mejor el funcionamiento de los mismos, y nos da una idea de como se aplican en la vida real.
Para finalizar. Agradecemos a los profesores por la paciencia y la dedicación que nos brindaron en este curso, y esperamos que este trabajo sea de su agrado.

Enlace al repositorio de GitHub:

https://github.com/jumpert/TareaFinal-PyEA.git

*Nota: Se recomienda clonar el repositorio para poder visualizar el proyecto completo. Para ello utiliza el comando: 'git clone https://github.com/jumpert/TareaFinal-PyEA.git'